Lab 6: Computing Gene Tree Probabilities with COAL

Laura Salter Kubatko
Departments of Statistics and
Evolution, Ecology, and Organismal Biology
The Ohio State University
lkubatko@stat.ohio-state.edu

May 12, 2010

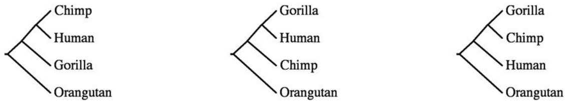
Example 1
Computing a Gene Tree Distribution

Example 2
Computing a Gene Tree Distribution

Example 3
Simulating Gene Trees Within a Species Tree

Example 4
Intraspecific Sampling

Example 1: Computing a Gene Tree Distribution



76.6%	11.4%	11.5%
79.1%	9.9%	9.9%

Observed proportions of each gene tree among ML phylogenies

Predicted proportions using parameters from Rannala & Yang, 2003.

COAL Example 1: Input Files

- How do we do this with COAL?
- Example files:
  - ex1_speciestree.tre:

```
(((H:1.0,C:1.0):1.2,G:2.2):4.2,O:6.4);
```

  - ex1_genetrees.tre:

```
(((H:1.0,C:1.0):1.0,G:2.0):1.0,O:3.0);
(((H:1.0,G:1.0):1.0,C:1.0):1.0,O:3.0);
(((C:1.0,G:1.0):1.0,H:1.0):1.0,O:3.0);
```

COAL Example 1: Input Files

- infile_example1:

```
[Infile for Example 1: Computing a Gene Tree Distribution]

begin coal;

ntax = 4;
taxa names = H C G O;
gene tree file = ex1_genetrees.tre;
species tree file = ex1_speciestree.tre;
intra = no;
ngtrees = 3;
nstrees = 1;
blstyle = none; [branch lengths read from file]
logfile = ex1_logfile.log;
outfile = ex1_output.out / gtopo probs;

end;
```

COAL Example 1: Output Files

- Once these files are stored in same directory as COAL, run COAL by typing "coal" from the prompt
- Results will go in file ex1_output.out:

```
1   GT:(((H,C),G),O)   0.790712390916
2   GT:(((H,G),C),O)   0.098892600081
3   GT:(((C,G),H),O)   0.098892600081
```

COAL Example 1: More Options

- infile_example1 with histories option:

```
[Infile for Example 1: Computing a Gene Tree Distribution]

begin coal;

ntax = 4;
taxa names = H C G O;
gene tree file = ex1_genetrees.tre;
species tree file = ex1_speciestree.tre;
intra = no;
ngtrees = 3;
nstrees = 1;
blstyle = none; [branch lengths read from file]
logfile = ex1_logfile.log;
outfile = ex1_output.out / gtopo histories probs;

end;
```

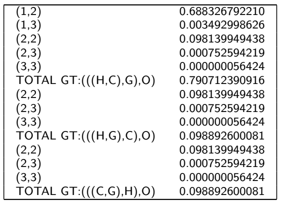
COAL Example 1: More Options

- Now the results file looks like this:

```
(1,2)                     0.688326792210
(1,3)                     0.003492998626
(2,2)                     0.098139949438
(2,3)                     0.000752594219
(3,3)                     0.000000056424
TOTAL GT:(((H,C),G),O)    0.790712390916
(2,2)                     0.098139949438
(2,3)                     0.000752594219
(3,3)                     0.000000056424
TOTAL GT:(((H,G),C),O)    0.098892600081
(2,2)                     0.098139949438
(2,3)                     0.000752594219
(3,3)                     0.000000056424
TOTAL GT:(((C,G),H),O)    0.098892600081
```

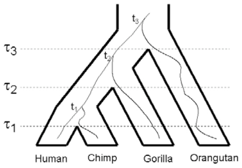
COAL Example 1: More Options

- Now the results file looks like this:



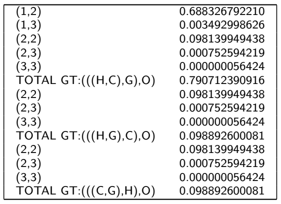
```
(1,2)                    0.688326792210
(1,3)                    0.003492998626
(2,2)                    0.098139949438
(2,3)                    0.000752594219
(3,3)                    0.000000056424
TOTAL GT:(((H,C),G),O)   0.790712390916
(2,2)                    0.098139949438
(2,3)                    0.000752594219
(3,3)                    0.000000056424
TOTAL GT:(((H,G),C),O)   0.098892600081
(2,2)                    0.098139949438
(2,3)                    0.000752594219
(3,3)                    0.000000056424
TOTAL GT:(((C,G),H),O)   0.098892600081
```

- $(1,2)$ means that the first coalescent event happens on branch 1 of the species tree (branches are labeled in a post-order traversal), and the second coalescent event happens on branch 2 of the species tree



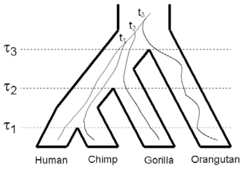
COAL Example 1: More Options

- Now the results file looks like this:



```
(1,2)                    0.688326792210
(1,3)                    0.003492998626
(2,2)                    0.098139949438
(2,3)                    0.000752594219
(3,3)                    0.000000056424
TOTAL GT:(((H,C),G),O)   0.790712390916
(2,2)                    0.098139949438
(2,3)                    0.000752594219
(3,3)                    0.000000056424
TOTAL GT:(((H,G),C),O)   0.098892600081
(2,2)                    0.098139949438
(2,3)                    0.000752594219
(3,3)                    0.000000056424
TOTAL GT:(((C,G),H),O)   0.098892600081
```

- $(3,3)$ indicates that both coalescent events happen above the root (which events these correspond to depends on the gene tree)



Example 2: Computing a Gene Tree Distribution

- In the previous example, we looked at the probabilities associated with three possible gene trees
- We might wish to completely characterize the gene tree distribution by computing the probabilities associated with *all* possible gene trees
- In addition to displaying values for a fixed set of species tree branch lengths, COAL will give us probabilities associated with both histories and gene trees which will allow us to characterize and study our distribution

Example 2: Computing a Gene Tree Distribution

- An example gene tree distribution - 4 taxa

COAL Example 2: Input Files

- Example files:
  - ex2_speciestree.tre:

```
(((A:1.0,B:1.0):1.0,C:2.0):1.0,D:3.0);
```

  - ex2_genetrees.tre:

```
(((A,D),B),C);
((A,(B,D)),C);
(((A,B),D),C);
((A,B),(C,D));
(((A,B),C),D);
(((A,D),C),B);
((A,(C,D)),B);
(((A,C),D),B);
((A,C),(B,D));
(((A,C),B),D);
(((B,D),C),A);
((B,(C,D)),A);
(((B,C),D),A);
((B,C),(A,D));
(((B,C),A),D);
```

COAL Example 2: Input Files

- infile_example2:

```
[Infile for Example 2: Computing a Gene Tree Distribution]

begin coal;

ntax = 4;
taxa names = A B C D;
gene tree file = ex2_genetrees.tre;
species tree file = ex2_speciestree.tre;
intra = no;
ngtrees = 15;
nstrees = 1;
blstyle = none; [branch lengths read from file]
logfile = ex2_logfile.log;
outfile = ex2_output.out / gtopo probs;

end;
```

COAL Example 2: Output Files

- Results will be in ex2_output.out

```
1     GT:(((A,D),B),C)     0.001017535494
2     GT:((A,(B,D)),C)     0.001017535494
3     GT:(((A,B),D),C)     0.098035528863
4     GT:((A,B),(C,D))     0.099053064357
5     GT:(((A,B),C),D)     0.555623375011
6     GT:(((A,D),C),B)     0.001017535494
7     GT:((A,(C,D)),B)     0.001017535494
8     GT:(((A,C),D),B)     0.020520809552
9     GT:((A,C),(B,D))     0.021538345046
10    GT:(((A,C),B),D)     0.078532254805
11    GT:(((B,D),C),A)     0.001017535494
12    GT:((B,(C,D)),A)     0.001017535494
13    GT:(((B,C),D),A)     0.020520809552
14    GT:((B,C),(A,D))     0.021538345046
15    GT:(((B,C),A),D)     0.078532254805
```

COAL Example 2: More Options

- infile_example2:

```
[Infile for Example 2: Computing a Gene Tree Distribution]

begin coal;

ntax = 4;
taxa names = A B C D;
gene tree file = ex2_genetrees.tre;
species tree file = ex2_speciestree.tre;
intra = no;
ngtrees = 15;
nstrees = 1;
blstyle = none; [branch lengths read from file]
logfile = ex2_logfile.log;
outfile = ex2_output.out / gtopo histories formulas probs.6;

end;
```

COAL Example 2: More Options

- New results - ex2_output.out

```
(3,3)     (1/18)p_{2 2}(T1)p_{3 3}(T2)          0.001018
TOTAL     GT:(((A,D),B),C)                      0.001018
(3,3)     (1/18)p_{2 2}(T1)p_{3 3}(T2)          0.001018
TOTAL     GT:((A,(B,D)),C)                      0.001018
(1,3)     (1/3)p_{2 1}(T1)p_{2 2}(T2)           0.077515
(2,3)     (1/3)p_{2 2}(T1)(1/3)p_{3 2}(T2)      0.019503
(3,3)     (1/18)p_{2 2}(T1)p_{3 3}(T2)          0.001018
TOTAL     GT:(((A,B),D),C)                      0.098036
(1,3)     (1/3)p_{2 1}(T1)p_{2 2}(T2)           0.077515
(2,3)     (1/3)p_{2 2}(T1)(1/3)p_{3 2}(T2)      0.019503
(3,3)     (2/18)p_{2 2}(T1)p_{3 3}(T2)          0.002035
TOTAL     GT:((A,B),(C,D))                      0.099053
(1,2)     (1/1)p_{2 1}(T1)p_{2 1}(T2)           0.399576
(1,3)     (1/3)p_{2 1}(T1)p_{2 2}(T2)           0.077515
(2,2)     (1/1)p_{2 2}(T1)(1/3)p_{3 1}(T2)      0.058011
(2,3)     (1/3)p_{2 2}(T1)(1/3)p_{3 2}(T2)      0.019503
(3,3)     (1/18)p_{2 2}(T1)p_{3 3}(T2)          0.001018
TOTAL     GT:(((A,B),C),D)                      0.555623
....
```

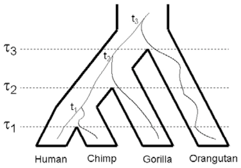
COAL Example 2: Interpreting Gene Tree Probability Formulas

- Let's look at the results for one of these gene trees, say the gene tree that matches the species tree:

```
(1,2)     (1/1)p_{2 1}(T1)p_{2 1}(T2)           0.399576
(1,3)     (1/3)p_{2 1}(T1)p_{2 2}(T2)           0.077515
(2,2)     (1/1)p_{2 2}(T1)(1/3)p_{3 1}(T2)      0.058011
(2,3)     (1/3)p_{2 2}(T1)(1/3)p_{3 2}(T2)      0.019503
(3,3)     (1/18)p_{2 2}(T1)p_{3 3}(T2)          0.001018
TOTAL     GT:(((A,B),C),D)                      0.555623
```

- The formulas tell us the probability of each history; e.g., for history (1,2), we have:

$$P_{21}(T1)P_{21}(T2)$$
$$= (1 - e^{-T1})(1 - e^{-T2})$$



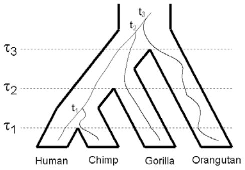
COAL Example 2: Interpreting Gene Tree Probability Formulas

- Let's look at the results for one of these gene trees, say the gene tree that matches the species tree:

```
(1,2)     (1/1)p_{2 1}(T1)p_{2 1}(T2)           0.399576
(1,3)     (1/3)p_{2 1}(T1)p_{2 2}(T2)           0.077515
(2,2)     (1/1)p_{2 2}(T1)(1/3)p_{3 1}(T2)      0.058011
(2,3)     (1/3)p_{2 2}(T1)(1/3)p_{3 2}(T2)      0.019503
(3,3)     (1/18)p_{2 2}(T1)p_{3 3}(T2)          0.001018
TOTAL     GT:(((A,B),C),D)                      0.555623
```

- For history (1,3), we have:

$$\tfrac{1}{3}P_{21}(T1)P_{22}(T2)$$
$$= \tfrac{1}{3}(1 - e^{-T1})(e^{-T2})$$



COAL Example 2: Interpreting Gene Tree Probability Formulas

- To get the probabilities for each of the 15 gene trees, we add the probabilities across histories:

History	Probability Expression	Probability
(1,2)	$P_{21}(T1)P_{21}(T2)$	$(1 - e^{-T1})(1 - e^{-T2})$
(1,3)	$\frac{1}{3}P_{21}(T1)P_{22}(T2)$	$(1 - e^{-T1})(e^{-T2})$
(2,2)	$\frac{1}{3}P_{22}(T1)P_{31}(T2)$	$(e^{-T2})(1 - \frac{3}{2}e^{-T2} + \frac{1}{2}e^{-3T2})$
(2,3)	$\frac{1}{9}P_{22}(T1)P_{32}(T2)$	$(e^{-T2})(\frac{3}{2}e^{-T2} - \frac{3}{2}e^{-3T2})$
(3,3)	$\frac{1}{18}P_{22}(T1)P_{33}(T2)$	$(e^{-T1})(e^{-3T2})$

- The probability of the tree (((A,B),C),D) is the sum of the expressions in the last column of the table above.

Example 2: Computing a Gene Tree Distribution

- Some notes:
  - This may take significantly longer than computing likelihoods for a single tree, and will depend on the number of taxa in several ways: e.g., more possible histories, more valid histories, more possible gene trees
  - With the COAL download, files containing all gene trees for up to 8 taxa are obtained (t4all, t5all, t6all, t7all, t8all)
  - A program to generate all possible gene trees is also packaged with the download (enum.c)
  - THINK carefully before you actually do this for more than 8 or so taxa .... there are many, many gene trees

Example 3: Simulating Gene Trees Within a Species Tree

- COAL can also be used to generate a sample of gene trees for a fixed species tree
- The infile is infile_example3:

```
[Infile for Example 3: Obtaining a Random Sample of Gene Trees]

begin coal;

ntax = 4;
taxa names = A B C D;
gene tree file = simulated;
species tree file = ex3_speciestree.tre;
intra = no;
theta = 2;
ngtrees = 100;
nstrees = 1;
blstyle = none; [branch lengths read from file]
logfile = ex3_logfile.log;
outfile = ex3_output.out;
seed1=12345;
seed2=67890;

end;
```

COAL Example 3: Output

- This will produce a sample of 100 gene trees from the gene tree distribution corresponding to the species tree in the file ex3_speciestree.tre
- The sampled gene trees will be stored in a file called simtrees.dat:

```
(((A:1.0294,B:1.0294):2.12647,C:3.15587):0.138865,D:3.29474);
(((A:1.00434,B:1.00434):1.09363,C:2.09797):1.09494,D:3.19291);
(((A:2.14825,B:2.14825):0.858405,D:3.00666):0.449066,C:3.45572);
(((A:1.02966,B:1.02966):2.05876,C:3.08842):0.0811206,D:3.16954);
(((A:1.19457,B:1.19457):1.0181,C:2.21267):3.10166,D:5.31433);
```

- Branch lengths in the gene trees are given in coalescent units: $\frac{t}{2N_e}$
- To convert to mutation units, branch lengths should be multiplied by $\frac{\theta}{2}$, where $\theta = 4N_e\mu$

Example 3: Utility of Simulation Component of COAL

COAL can be used to simulate sequence data under the coalescent model:

1. Simulate gene trees in COAL: command line version available to use in scripts
   - coal i=infile_example3 seed1=6372 seed2=8493 ngtrees=5
   - include any options that differ from current infile

Example 3: Utility of Simulation Component of COAL

COAL can be used to simulate sequence data under the coalescent model:

1. Simulate gene trees in COAL: command line version available to use in scripts
   - coal i=infile_example3 seed1=6372 seed2=8493 ngtrees=5
   - include any options that differ from current infile
2. Run a sequence data simulator like Seq-Gen
   - awk '{print "[300]" $1}' simtrees.dat > simtrees2.dat
   - seq-gen -mHKY -t1.0 -f0.25 0.25 0.25 0.25 -l1500 -p5 -z20584 -on -s0.005 < simtrees2.dat > my_sequence_data.nex
   - This will generate a concatenated data set with five partitions of length 300 nucleotides each from the gene trees in the file simtrees.dat under the JC69 model with $\theta = 0.01$. The data would be stored in the file my_sequence_data.nex.

Example 3: Utility of Simulation Component of COAL

- Notice that trees are written in a particular order - easy to parse file to pull off branch lengths, or to search for topology of interest
- Can simulate very large samples of trees without crashing - several hundred thousand
- Simulation of intraspecific samples is not yet incorporated - coming soon!

Example 4: Intraspecific Sampling

- COAL can also compute gene tree probabilities in the case of more than one sample per species
- Consider the case where two samples are taken from taxa A and B, and we interested in the probability of the gene tree for which the within taxon samples are monophyletic:

```
((((A-1:1.0,A-2:1.0):1.0,(B-1:1.0,B-2:1.0):1.0):1.0,C:3.0):1.0,D:4.0);
```

- Note: A specific naming convention is required for the within-taxon samples.

Example 4: Intraspecific Sampling

- The infile is infile_example4:

```
[Infile for Example 4: Intraspecific Sampling]

begin coal;
ntax = 4 | nstaxa 4;
taxa names = A B C D;
ngtaxa 2 2 1 1;
gene tree file = ex4_genetrees.tre;
species tree file = ex4_speciestree.tre;
intra = yes;
ngtrees = 1;
nstrees = 1;
blstyle = none; [branch lengths read from file]
logfile = ex4_logfile.log;
outfile = ex4_output.out / prob;

end;
```